CSC 446, HW#4, Kefu Zhu

1. Bishop 4.8

Using (4.57) and (4.58), derive the result (4.65) for the posterior class probability in the two-class generative model with Gaussian densities, and verify the results (4.66) and (4.67) for the parameters w and w_0.

(4.57)

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1)+p(x|C_2)p(C_2)} = \frac{1}{1+\exp(-a)} = \sigma(a)$$

(4.58)

$$a = \ln\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$$

(4.65)

$$p(C_1|x) = \sigma(\mathrm{w}^T x + w_0)$$

(4.66)

$$\mathrm{w} = \sum^{-1}(\mu_1 - \mu_2)$$

(4.67)

$$w_0 = -\frac{1}{2}\mu_1^T\sum^{-1}\mu_1 + \frac{1}{2}\mu_2^T\sum^{-1}\mu_2 + \ln\frac{p(C_1)}{p(C_2)}$$

Answer:

Because the conditional desnities are Gaussian, we then have

$$p(x|C_k) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\sum|^{1/2}}\exp\{-\frac{1}{2}(x-\mu_k)^T\sum^{-1}(x-\mu_k)\}$$

$\therefore a = \ln\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$

$= \ln\frac{p(x|C_1)}{p(x|C_2)} + \ln\frac{p(C_1)}{p(C_2)}$

$= \ln\frac{\exp\{-\frac{1}{2}(x-\mu_1)^T\sum^{-1}(x-\mu_1)\}}{\exp\{-\frac{1}{2}(x-\mu_2)^T\sum^{-1}(x-\mu_2)\}} + \ln\frac{p(C_1)}{p(C_2)}$

$= -\frac{1}{2}(x-\mu_1)^T\sum^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\sum^{-1}(x-\mu_2) + \ln\frac{p(C_1)}{p(C_2)}$

$= -\frac{1}{2}(x^T\sum^{-1}x - x^T\sum^{-1}\mu_1 - \mu_1^T\sum^{-1}x + \mu_1^T\sum^{-1}\mu_1) + \frac{1}{2}(x^T\sum^{-1}x - x^T\sum^{-1}\mu_2 - \mu_2^T\sum^{-1}x + \mu_2^T\sum^{-1}\mu_2) + \ln\frac{p(C_1)}{p(C_2)}$

$= (-\frac{1}{2}x^T \sum^{-1} x + \frac{1}{2}x^T \sum^{-1} x) + (\mu_1^T \sum^{-1} x - \mu_2^T \sum^{-1} x) + (-\frac{1}{2}\mu_1^T \sum^{-1} \mu_1 + \frac{1}{2}\mu_2^T \sum^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)})$

$= (\mu_1^T \sum^{-1} x - \mu_2^T \sum^{-1} x) + (-\frac{1}{2}\mu_1^T \sum^{-1} \mu_1 + \frac{1}{2}\mu_2^T \sum^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)})$

$\because \mu_1, \mu_2$ are $(n \times 1)$ vectors. Since x is also a $(n \times 1)$ vector, $\sum$ is a symmetrically diagonal matrix.

$\therefore \mu_1^T \sum^{-1} x$ is a scalar $\rightarrow \mu_1^T \sum^{-1} x = x^T \sum^{-1} \mu_1, \ \mu_2^T \sum^{-1} x = x^T \sum^{-1} \mu_2$

Hence, $a = \sum^{-1}(\mu_1 - \mu_2)^T x + (-\frac{1}{2}\mu_1^T \sum^{-1} \mu_1 + \frac{1}{2}\mu_2^T \sum^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)})$, which is in the form of $a = \mathrm{w}^T x + w_0$, where

$$\begin{cases} \mathrm{w} = \sum^{-1}(\mu_1 - \mu_2) \\ w_0 = -\frac{1}{2}\mu_1^T \sum^{-1} \mu_1 + \frac{1}{2}\mu_2^T \sum^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)} \end{cases}$$

2. Bishop 5.4

Consider a binary classification problem in which the target values are $t \in \{0, 1\}$, with a network output $y(x, w)$ that represents $p(t = 1|x)$, and suppose that there is a probability ϵ that the class label on a training data point has been incorrectly set. Assuming independent and identically distributed data, write down the error function corresponding to the negative log likelihood. Verify that the error function (5.21) is obtained when $\epsilon = 0$. Note that this error function makes the model robust to incorrectly labelled data, in contrast to the usual error function.

(5.21)

$$E(w) = -\sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

Answer:

Consider a single output t_n for input x_n, we have

$p(t_n = 1|x_n) = p(correctly\ labeled) \cdot p(t_n = 1|x_n) + p(incorrectly\ labeled) \cdot p(t_n = 0|x_n)$

$= (1 - \epsilon) \cdot y_n(x_n, w) + \epsilon \cdot (1 - y_n(x_n, w))$

Because the conditional distribution of targets given inputs is a Bernoulli distribution, so we have

$p(t|x) = p(t = 1|x)^t \cdot p(t = 0|x)^{1-t}$

Then our negative log likelihood loss function can be expanded as following

$E(w) = -\sum_{n=1}^N \ln p(t_n|x_n)$

$= -\sum_{n=1}^N \ln[\ p(t_n = 1|x_n)^{t_n} \cdot p(t_n = 0|x_n)^{1-t_n}\ ]$

$= -\sum_{n=1}^N t_n \ln p(t_n = 1|x_n) + (1 - t_n) \ln p(t_n = 0|x_n)$

$= -\sum_{n=1}^N t_n \ln[\ (1 - \epsilon) \cdot y_n(x_n, w) + \epsilon \cdot (1 - y_n(x_n, w))\ ] + (1 - t_n) \ln[\ (1 - (1 - \epsilon)) \cdot y_n(x_n, w) - \epsilon \cdot (1 - y_n(x_n, w))\ ]$

Consider the case where $\epsilon = 0$

$E(w) = -\sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$

3. Bishop 5.7

Show the derivative of the error function (5.24) with respect to the activation a_k for output units having a softmax activation function satisfies (5.18).

(5.24)

$$E(w) = -\sum_{n=1}^N \sum_{k=1}^K t_{kn} \ln y_k(x_n, w)$$

(5.18)

$$\frac{\partial E}{\partial a_k} = y_k - t_k$$

Answer:

Denote The softmax activation function as $\sigma(a_j) = \frac{e^{a_j}}{\sum_{k=1}^K e^{a_k}}$ for $j = 1, 2, \ldots, K$

By the chain rule, $\frac{\partial E}{\partial a_k} = \frac{\partial E}{\partial y_i} \cdot \frac{\partial y_i}{\partial a_k}$

- $\frac{\partial E}{\partial y_i} = -\sum_{n=1}^N \sum_{i=1}^K \frac{t_{in}}{y_i}$
- $\frac{\partial y_i}{\partial a_k}$
  - $i = k$

    By the quotient rule, $\frac{\partial y_i}{\partial a_k} = \frac{\partial y_k}{\partial a_k} = \frac{e^{a_k} \sum_{k=1}^K e^{a_k} - e^{a_k} e^{a_k}}{(\sum_{k=1}^K e^{a_k})^2} = \frac{e^{a_k}}{\sum_{k=1}^K e^{a_k}} - (\frac{e^{a_k}}{\sum_{k=1}^K e^{a_k}})^2 = y_k(1 - y_k)$
  - $i \neq k$

    $\frac{\partial y_i}{\partial a_k} = \frac{0 - e^{a_k} e^{a_i}}{(\sum_{k=1}^K e^{a_k})^2} = -y_k y_i$

$\therefore \frac{\partial E}{\partial a_k} = -\sum_{n=1}^N \sum_{i=1}^{k-1} \frac{t_{in}}{y_i} \cdot (-y_k y_i) - \sum_{n=1}^N \frac{t_{in}}{y_k} \cdot y_k(1 - y_k) - \sum_{n=1}^N \sum_{i=k+1}^K \frac{t_{in}}{y_i} \cdot (-y_k y_i)$

$= -\sum_{n=1}^N [\sum_{i=1}^{k-1} \frac{t_{in}}{y_i} \cdot (-y_k y_i) + \sum_{i=k+1}^K \frac{t_{in}}{y_i} \cdot (-y_k y_i) + t_{in} \cdot (1 - y_k)]$

$= -\sum_{n=1}^N y_{kn}(-\sum_{i=1}^{k-1} t_{in} + \frac{t_{kn}}{y_{kn}})$

$\because$ For $n = 1, 2, \ldots, N$, $\sum_{i=1}^K t_{in} = 1$

$\therefore \frac{\partial E}{\partial a_k} = -\sum_{n=1}^N y_{kn}(-1 + \frac{t_{kn}}{y_{kn}}) = \sum_{n=1}^N y_{kn} - t_{kn}$

4. Bishop 5.17

Consider a squared loss function of the form

(5.193)

$$E = \frac{1}{2} \int \int \{y(x, w) - t\}^2 \, p(x, t) \, \mathrm{d}x\mathrm{d}t$$

where $y(x, w)$ is a parametric function such as a neural network. The result (1.89)

(1.89)

$$y(x) = \frac{\int tp(x,t)\mathrm{d}t}{p(x)} = \int tp(t|x)\mathrm{d}t = E_t[t|x]$$

shows that the function $y(x, w)$ that minimizes this error is given by the conditional expectation of t given x. Use this result to show that the second derivative of E with respect to two elements w_r and w_s of the vector w, is given by

(5.194)

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \int \frac{\partial y}{\partial w_r} \frac{\partial y}{\partial w_s} p(x)\mathrm{d}x$$

Note that, for a finite sample from $p(x)$, we obtain (5.84).

Answer:

$\frac{\partial^2 E}{\partial w_r \partial w_s} = \frac{1}{2} \frac{\partial}{\partial w_r} \int \int 2\{y(x, w) - t\} \frac{\partial y(x,w)}{\partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t$

$= \int \int (\frac{x^2}{2})[\frac{\partial y(x,w)}{\partial w_r} \cdot \frac{\partial y(x,w)}{\partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t + \{y(x, w) - t\} \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t]$

$= \int \frac{\partial y}{\partial w_r} \cdot \frac{\partial y}{\partial w_s} \, p(x)\mathrm{d}x + \int \int \{y(x, w) - t\} \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t$

where $\int \int \{y(x, w) - t\} \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t$

$= \int \int y(x, w) \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t - \int \int t \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t$

$= \int \int y(x, w) \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t - \int \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \int t \, p(x, t) \, \mathrm{d}x\mathrm{d}t$

As stated in the question, based on (1.89)

$y(x) = \frac{\int tp(x,t)\mathrm{d}t}{p(x)} \rightarrow \int tp(x, t)\mathrm{d}t = y(x)p(x)$

Hence,

$\int \int \{y(x, w) - t\} \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t$

$= \int \int y(x, w) \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \, p(x, t) \, \mathrm{d}x\mathrm{d}t - \int \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \int t \, p(x, t) \, \mathrm{d}x\mathrm{d}t$

$$= \int \int y(x,w) \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} \, p(x) \, \mathrm{d}x - \int \frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} y(x,w) p(x) \, \mathrm{d}x$$

$$= 0$$

Taking this back to the original equation, we can derive that

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \int \frac{\partial y}{\partial w_r} \cdot \frac{\partial y}{\partial w_s} \, p(x) \mathrm{d}x + 0$$